




Custom Stop Words in R

Data Files: (1) file.csv (2) csw.csv



Data

> CSW				
Name	Date modified	Type	Size	
 csw	18-08-2020 16:29	Microsoft Excel Co...	1 KB	
 csw	18-08-2020 17:10	R File	1 KB	
 file	18-08-2020 16:41	Text Document	1 KB	

	A	B
1	not	
2	at	
3	all	
4	difficult	
5		

My Data

file - Notepad

File Edit Format View Help

Learning R is not at ALL
difficult, it's amazingly easy
and fun.

Ln 1, Col 1 100% Windows (CRLF) UTF-8

My Stop
Words

Set Working Directory

```
1 # Jesus is Great!  
2 setwd("C:/Users/Dr Vinod/Desktop/csw")  
3 getwd()  
4 library(tm)
```

```
> getwd()  
[1] "C:/Users/Dr Vinod/Desktop/csw"
```



Read Data

```
6 data <- readLines("file.txt")  
7 data
```

```
> data <- readLines("file.txt")  
Warning message:  
In readLines("file.txt") : incomplete final line found on 'file.txt'  
> data  
[1] "Learning R is not at ALL difficult, it's amazingly easy and fun."
```



Put it in a vector



```
8 text <- VectorSource(data)
9 text
```

```
> text <- VectorSource(data)
> text
$`encoding`
[1] ""

$length
[1] 1

$position
[1] 0

$reader
function (elem, language, id)
{
  if (!is.null(elem$uri))
    id <- basename(elem$uri)
  PlainTextDocument(elem$content, id = id, language = language)
}
<bytecode: 0x00000000c427950>
<environment: namespace:tm>

$content
[1] "Learning R is not at ALL difficult, it's amazingly easy and fun."

attr(,"class")
[1] "VectorSource" "simpleSource" "Source"
```

Create Corpus

```
10 text1 <- VCorpus(text)
11 text1
```

```
> text1 <- VCorpus(text)
> text1
<<VCorpus>>
Metadata: corpus specific: 0, document level (indexed): 0
Content: documents: 1
```



```
> csw <- read.csv("csw.csv", header = FALSE)
> csw # as data frame
  V1
1 not
2 at
3 all
4 difficult
> class(csw) # data frame
[1] "data.frame"
> csw1 <- as.character(csw$V1)
> csw1 # char (in a row)
[1] "not"      "at"      "all"      "difficult"
> class(csw1) # char
[1] "character"
```

[illegible]

```
> mySW <- c(csw1, stopwords())
> mySW # first 4 are custom words
[1] "not"      "at"       "all"      "difficult" "i"
[6] "me"       "my"       "myself"   "we"        "our"
[11] "ours"     "ourselves" "you"      "your"      "yours"
[16] "yourself" "yourselves" "he"       "him"       "his"
```

```
[161] "both"      "each"      "few"      "more"      "most"
[166] "other"     "some"     "such"     "no"       "nor"
[171] "not"      "only"     "own"      "same"     "so"
[176] "than"     "too"      "very"
```


Observe L, R & ALL \rightarrow l, r as all

```
21 #_____
22 text2 <- tm_map(text1, content_transformer(tolower))
23 text2[[1]]$content
24
```

```
> text2 <- tm_map(text1, content_transformer(tolower))
> text2[[1]]$content
[1] "learning r is not at all difficult, it's amazingly easy and fun."
```



Stop Words, removed!

```
24  
25 text3 <- tm_map(text2, removeWords, mySw)  
26 text3[[1]]$content  
27
```

```
> text3 <- tm_map(text2, removewords, mySw)
> text3[[1]]$content
[1] "learning r      amazingly easy fun."
```



Why gap?

You got what you wanted!

```
27  
28 text4 <- tm_map(text3, stripwhitespace)  
29 text4[[1]]$content
```

```
> text4 <- tm_map(text3, stripwhitespace)  
> text4[[1]]$content  
[1] "learning r , amazingly easy fun."
```



